Genome-wide association study of INDELs identified four novel susceptibility loci associated with lung cancer risk

Juncheng Dai,[1,2] † Mingtao Huang,[1] † Christopher I. Amos,[3] Rayjean J. Hung,[4] Adonina Tardon,[5] Angeline Andrew,[6] Chu Chen,[7] David C. Christiani,[8] Demetrius Albanes,[9] Gadi Rennert,[10] Jingyi Fan,[1] Gary Goodman,[11] Geoffrey Liu,[12] John K. Field,[13] Kjell Grankvist,[14] Lambertus A. Kiemeney,[15] Loic Le Marchand,[16] Matthew B. Schabath,[17] Mattias Johansson,[18] Melinda C. Aldrich,[19] Mikael Johansson,[20] Neil Caporaso,[9] Philip Lazarus,[21] Stephan Lam,[22] Stig E. Bojesen,[23,24] Susanne Arnold,[25] Maria Teresa Landi,[9] Angela Risch,[26] H-Erich Wichmann,[27] Heike Bickeboller,[28] Paul Brennan,[29] Sanjay Shete,[30] Olle Melander,[31] Hans Brunnstrom,[31] Shan Zienolddiny,[32] Penella Woll,[33] Victoria Stevens,[34] Zhibin Hu,[1,2] Hongbing Shen,[1,2] *

1 Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China

2 Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Nanjing Medical University, Nanjing 211166, China

3 Department of Medicine, Epidemiology Section, Institute for Clinical and Translational Research, Baylor Medical College, Houston, Texas, 77030, USA

4 Epidemiology Division, Lunenfeld-Tanenbuaum Research Institute, Sinai Health System, Toronto, Ontario, M5T 3L9, Canada

5 Faculty of Medicine, University of Oviedo and CIBERESP, Oviedo, 33006, Spain

6 Department of Neurology, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire, 3756, USA

7 Department of Epidemiology, Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109-1024, USA

8 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, 2115, USA

9 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, 20892-7150, USA

10 Technion Faculty of Medicine, Carmel Medical Center, Israel Institute of Technology, Haifa, Israel

11 Public Health Sciences Division, Swedish Cancer Institute, Seattle, Washington,

98109, USA
12 Epidemiology Division, Princess Margaret Cancer Center, Toronto, Ontario, M5G 2M9, Canada
13 Roy Castle Lung Cancer Research Programme, Department of Molecular & Clinical Cancer Medicine, University of Liverpool, Liverpool, L69 3BX, UK
14 Unit of Clinical Chemistry, Department of Medical Biosciences, Umeå University, Umeå, 901 85, Sweden
15 Department of Health Evidence, Radboud university medical center, Nijmegen, Germany
16 Department of Epidemiology, University of Hawaii Cancer Center, Honolulu, Hawai'I, 96813, USA
17 Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, 33612-9497, USA
18 Genetic Epidemology Group, International Agency for Research on Cancer, Lyon, 69372 CEDEX 08, France
19 Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center, Nashville, Tennessee, 37232, USA
20 Department of Radiation Sciences, Umeå University, Umeå, 901 85, Sweden
21 Washington State University College of Pharmacy, Spokane, Washington, 99210-1495, USA
22 Department of Integrative Oncology, British Columbia Cancer Agency, Vancouver, British Columbia, V5Z 1L3, Canada
23 Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, 2200, Denmark
24 Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 2200, Denmark
25 Markey Cancer Center, University of Kentucky, Lexington, Kentucky, 40508, USA
26 Cancer Center Cluster Salzburg at PLUS, Department of Molecular Biology, University of Salzburg, Heidelberg, 5020, Austria
27 Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig Maximilians University, Munich, Bavaria, Germany
28 Department of Genetic Epidemiology, University Medical Center Goettingen, Goettingen, 37073, Germany

29 Genetic Epidemology Group, International Agency for Research on Cancer, Lyon, 69372 CEDEX 08, France

30 Department of Epidemiology, University of Texas, MD Anderson Cancer Center, Houston, Texas, 77030, USA

31 Clinical Sciences, Lund University, Lund, 22100, Sweden

32 National Institute of Occupational Health (STAMI), Oslo, Norway

33 Academic Unit of Clinical Oncology, University of Sheffield, Sheffield, S10 25J, UK

34 Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, 30303, USA

† These authors contributed equally to this work.

* Correspondence to: Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China

Tel (fax): +86-25-868-68439, E-mail: hbshen@njmu.edu.cn

Abstract

Genome-wide association studies (GWAS) have identified 45 single nucleotide polymorphisms (SNPs) associated with lung cancer. Only less than SNPs, small insertions and deletions (INDELs) are the second most abundant genetic polymorphisms in the human genome. INDELs are highly associated with multiple human diseases, including lung cancer. However, limited studies with large-scale samples have been available to systematically evaluate the effects of INDELs on lung cancer risk. Here, we performed a large-scale meta-analysis to evaluate INDELs and their risk for lung cancer in 23,202 cases and 19,048 controls. Functional annotations were performed to further explore the potential function of lung cancer risk INDELs. Conditional analysis was used to clarify the relationship between INDELs and SNPs. Four new risk loci were identified in genome-wide INDEL analysis (1p13.2: rs5777156, Insertion, OR = 0.92, $P = 9.10\times10^{-8}$; 4q28.2: rs58404727, Deletion, OR = 1.19, $P = 5.25\times10^{-7}$; 12p13.31: rs71450133, Deletion, OR = 1.09, $P = 8.83\times10^{-7}$; and 14q22.3: rs34057993, Deletion, OR = 0.90, $P = 7.64\times10^{-8}$). The eQTL analysis and functional annotation suggested that INDELs might affect lung cancer susceptibility by regulating the expression of target genes. After conducting conditional analysis on potential causal SNPs, the INDELs in the new loci were still nominally significant. Our findings indicate that INDELs could be potentially functional genetic variants for lung cancer risk. Further functional experiments are needed to better understand INDEL mechanisms in carcinogenesis.

Keywords

INDELs; Lung cancer; Genome-wide association studies (GWAS)

Introduction

Lung cancer is one of the most frequently diagnosed cancers and the leading cause of cancer mortality worldwide [1]. It is estimated that more than 1.8 million new lung cancer cases occurred in 2012, accounting for approximately 13% of total cancer diagnoses [2]. Although tobacco smoking is a major lung cancer risk factor, genetic factors also play an important role in lung carcinogenesis. According to previous studies, common SNPs can explain approximately 12% ~ 21% heritability in lung cancer in Asian and European populations [3,4]. Genome-wide association studies (GWAS) have previously identified 45 susceptibility loci associated with lung cancer [5], and single nucleotide polymorphisms (SNPs) in the *CHRNA3*, *CHRNA5*, *TERT* and human leukocyte antigen (HLA) regions showed consistent and robust associations in different studies.

To date, the vast majority of studies have focused on the relationship between SNPs and lung cancer. Small insertions and deletions (INDELs), which are another type of variations, also play an important role in lung carcinogenesis. INDELs are defined as short insertions and deletions (ranging from 1 to 10,000 bp) in the human genome [6,7]. As important genetic variations, INDELs are the second most abundant genetic polymorphisms in the human genome, only less than SNPs [8]. The final phase of the 1000 Genomes Project has characterized more than 3.4 million INDELs in 88 million variant sites in the human genome, and compared with phase I, the number of INDELs increased by 70% [8]. This provides a comprehensive panel to explore the effects of INDELs. INDELs in the genome are highly associated with multiple human diseases; nearly 24% of Mendelian diseases are caused by INDELs based on the Human Gene Mutation Database (HGMD) [9]. Over the past decade, the development of high-throughput sequencing has made it possible to detect INDELs in individual genomes. Next-generation sequencing (NGS) analyses have identified INDELs across multiple cancer types [10,11]; however, these INDELs were at the somatic level with low frequency. At the germline level, INDELs have been described as associated with cancers in case-control studies by genotyping or genomic imputation. For example, a single INDEL in the 6q25.3 locus, which is

related to the *SLC22A1* and *SLC22A2* genes, increased the risk of prostate cancer in a multi-ethnic GWAS [12]. Another study in a Chinese population found that a 5-bp INDEL in the *GAS5* gene increased hepatocellular carcinoma risk [13]. For lung cancer risk, Sun T *et al*. reported a six-nucleotide deletion variant in the *CASP8* promoter was related with reduced risk of multiple cancers, including lung cancer [PMID: 17450141]. In addition, Liu G *et al.* found two insertion variant in *BRM* promoter region were also associated with the increased risk of lung cancer [PMID: 21478907]. However, limited studies with large-scale samples have been available to systematically evaluate the effects of INDELs on lung cancer risk. In this study, we aimed to investigate the relationship between INDELs and lung cancer risk at a genome-wide level. To accomplish this, we conducted a large-scale case-control study with 23,202 lung cancer cases and 19,048 controls to dissect the associations between INDELs and lung cancer risk among European and Asian populations.

Material and Methods

Study population

In this study, we integrated three published lung cancer GWAS, including the TRICL-ILCCO OncoArray European data (The OncoArray Consortium lung cancer GWAS: 43,398 participants in total, European population) [14], the DCEG Lung Cancer Study (the National Cancer Institute lung cancer GWAS: 5,716 cases and 5,821 controls, European population) [15], and our published NJMU GWAS data (Nanjing Medical University lung cancer GWAS from Nanjing and Beijing: 2,331 cases and 3,077 controls, Chinese population) [16]. Briefly, for the TRICL-ILCCO OncoArray data, we used the same quality control strategies in the previous paper [14]. The DCEG Lung Cancer Study was applied from the Genotypes and Phenotypes (dbGAP) database (Project ID: phs000336.v1.p1) [15]. Considering the duplication of samples within the TRICL-ILCCO OncoArray data, 3,251 samples were removed when IBD (identity-by-descent) > 0.45. Consequently, 2,427 cases and 1,944 controls in the DCEG Lung Cancer Study were kept for further analysis. For the NJMU GWAS data, standard sample quality control strategies were also performed according to the original paper [16]. Finally, a total of 23,202 cases and

19,048 controls were included for further analysis (**Table S1)**. Each study was approved by the local institutional review board.

Genotype quality control and imputation

The details of the imputation procedures used in the TRICL-ILCCO OncoArray project have been described previously [14,17]. Briefly, SHAPETIT V2 and IMPUTE2 were used for phasing and imputation, respectively. The 1000 Genomes Project Phase III database (released at October, 2014) was used as a reference dataset. After imputation, there were 1,857,403 INDELs in the TRICL-ILCCO OncoArray data. Then, we performed standard quality control on the imputed INDELs data by excluding the data with the following characteristics: (1) imputation quality INFO < 0.9; (2) genotyping call rate < 95%; (3) minor allele frequency (MAF) in controls < 0.01; or (4) Hardy-Weinberg equilibrium (HWE) $<1\times10^{-12}$ in cases or $<1\times10^{-7}$ in controls. We also excluded 17,812 INDELs located in genome segmental duplication regions [18], which may lead to inaccuracy during imputation. Thus, the total number of TRICL-ILCCO OncoArray INDELs was 694,395. For the DCEG GWAS and NJMU GWAS data, the imputation procedures have been previously described [19,20]. We conducted the same quality control criteria on the DCEG GWAS and NJMU GWAS imputation data. Finally, we obtained 484,196 overlapped INDELs for the subsequent analysis (**Figure S1**).

eQTL and differential expression analysis

We used the Genotype-Tissue Expression (GTEx) Project expression quantitative trait locus (eQTL) database (V7 release) for identified INDELs. We searched each INDEL-gene pair eQTL analysis result in lung tissue. Due to lack of information of INDEL rs71450133 in GTEx database, we use SNP rs28435996 which showed high LD ($r^2 = 0.94$) with rs71450133 as a tagging SNP. Differential expression analyses were performed using data from The Cancer Genome Atlas (TCGA) project [21,22]. A total of 106 paired lung tumor tissues and adjacent tissues from the TCGA database were used to performed differential expression analyses using Wilcoxon paired test.

In silico functional annotation and rank scoring system development

We combined multiple sources of public functional annotation databases to explore the potential function of the INDELs, similar strategy was also applied in the recent largest breast cancer GWAS study with the INQUISIT algorithm [PMID: 29059683]. Genomic regulatory region and functional score were used to evaluate INDELs and SNP showed high LD with them. Regulatory elements, including promoter, enhancer, and transcription factor binding sites (TFBS) data were based on the Encyclopedia of DNA Elements (ENCODE) Project A549 human lung cancer cell line data [23]. Four annotation database, including 3DSNP [24], Combined Annotation-Dependent Depletion (CADD) [25], Phenotype-Informed Noncoding Element Scoring (PINES) [26] and RegulomeDB [27] were also used to identify the potential pathogenicity and function of the INDELs. We developed a rank scoring system to integrate all these data together and INDELs identified in this study, as well as SNPs which showed a high LD ($r2 > 0.6$) relationship with INDELs were all annotated by this rank scoring system.

We generated binary variables, feature rank, to represent importance of each variant in each database, 1 defined as more important and 0 defined as less important. For chromatin biofeatures data, as mentioned above, promoter, enhancer and TFBS, if INDELs or SNPs located in the regulatory region, the feature rank were defined as 1, else as 0. For four annotation databases (3DSNP, CADD, PINES and RegulomeDB) with scores, if a variant's score in the top 10% of corresponding INDEL LD block, the feature rank was defined as 1 (more important), otherwise it was defined as 0 (less important). Finally, all feature ranks of seven annotations were accumulated as a final score for each INDEL and SNP, ranging from 0 to 7. The variant with the highest score was considered as a potentially causal variant.

Statistical analysis

For the three GWAS studies, the association testing for each INDEL was performed using the SNPTEST (v2.5.4) software, which is based on a probabilistic dosage model adjusting for age, gender, and the first three principal components in

the TRICL-ILCCO OncoArray; age, gender, and the first principal component in the DCEG GWAS; and age, gender, pack-years, and the first principal component in the NJMU GWAS. Meta-analysis (fixed-effect model) was conducted to combine individual association estimates from the three GWAS datasets. Testing for differences in the genetic effects across the three studies was assessed by using the I^2 and P values calculated from Cochran's Q statistic. Meta-analysis was conducted using the GWAMA software. Subgroup analysis was performed for baseline characteristics, including age, gender, histology, and smoke status. For the conditional analysis, a multivariate logistic regression model adjusting for age, gender, the first three principal components and known lung cancer risk variants was used with the TRICL-ILCCO OncoArray.

General analyses were performed using the R software (version 3.3.1). $P \leq 0.05$ was used as the threshold of statistical significance and all statistical tests were two-sided. A suggestive threshold of 1.0×10^{-6} was used to present significant INDELs [PMID: 19915574, 23722424], bonferroni correction was also applied to account for multiple comparisons (threshold: $0.05/484{,}196 = 1.03\times10^{-7}$).

Results

Study overview

In this study, we imputed a total of 484,196 INDELs based on 23,202 lung cancer cases and 19,048 controls. Nineteen INDELs along with 11 loci were identified as being significantly associated with lung cancer risk at a suggestive threshold of 1.0×10^{-6} (**Figure 1**; **Table 1**; **Table 2**). Among them, four loci (1p13.2, 4q28.2, 12p13.31 and 14q22.3) were novel risk loci for lung cancer, while seven of them have been previously reported as lung cancer risk loci as indicated by SNPs (5p15.33, 6p21.32, 6p21.33, 6p22.1, 6p22.2, 11q23.3 and 15q25.1). The results of INDELs in three studies were listed in **Table S2**.

Four new risk loci were identified in our genome-wide INDEL analysis (**Table 1**), including rs5777156 in 1p13.2 (Insertion, OR = 0.92, 95%CI = 0.89-0.95, $P = 9.10\times10^{-8}$); rs58404727 in 4q28.2 (Deletion, OR = 1.19, 95%CI = 1.11-1.28, $P = 5.25\times10^{-7}$); rs71450133 in 12p13.31 (Deletion, OR = 1.09, 95%CI = 1.05-1.13, P =

8.83×10^{-7}); and rs34057993 in 14q22.3 (Deletion, OR = 0.90, 95%CI = 0.87-0.94, $P = 7.64\times10^{-8}$). INDELs rs5777156 and rs34057993 were still significant after Bonferroni correction ($P < 1.03\times10^{-7}$). There was no evidence of heterogeneity among the studies for the new risk loci. Subgroup analyses on the four new INDELs from the OncoArray data are summarized in **Table S3**. No evidence of heterogeneity was observed for the new risk loci among age, gender, smoking status, histology type and ethnicity, which implied the effects of the new risk loci were robust.

INDELs in known lung cancer risk loci

The results for 15 INDELs in known lung cancer risk loci are presented in **Table 2**. At 15q25.1, a well-known lung cancer susceptibility locus related to nicotine addiction, INDELs harbored the lowest *P* value (rs577626090, Deletion, OR = 1.29, 95%CI = 1.25-1.33, $P = 9.91\times10^{-64}$). INDELs also reached the significance threshold in 5p15.33 and the human leukocyte antigen (HLA) region. We validated the recently reported Oncoarray risk locus, which correlated with 11q23.3 in our analysis (rs139157129, Deletion, OR = 0.93, 95%CI =0.90-0.95, $P = 1.90\times10^{-7}$). INDELs in the known loci showed strong effects, and 10 of the 15 INDELs were still significant after Bonferroni correction (*P* threshold = 1.03×10^{-7}).

Functional annotations of new regions

Because the underlying mechanisms of known regions have been well illustrated, we performed functional annotations on the four new loci in this study. To explore the potential functions of the INDELs, we performed eQTL and differential expression analyses based on GTEx lung tissue data and TCGA lung cancer data for these four new regions. In GTEx lung eQTL database, we identified a total of 10 genes that showed significant cis-eQTL results (*P* value < 0.05), and 5 of them were related to cancer in previous studies. INDEL rs58404727 was a lung cis-eQTL for *HSPA4L*, which encodes heat shock protein family A (Hsp70) member 4 like. *HSPA4L* expression was significantly upregulated in lung tumor tissues compared with adjacent lung tissues ($P = 4.57\times10^{-13}$; **Figure 3**). For INDEL rs71450133, its

tag SNP rs28435996 was associated with decreased *GAPDH*, *TPI1*, *USP5* expression and increased *MLF2* expression. In the differential expression analysis, *GAPDH*, *TPI1*, *USP5* and *MLF2* were all significantly upregulated in lung tumor tissues compared with adjacent lung tissues (**Figure 3**). The full results from the cis-eQTL and differential expression analyses are presented in **Table S4**.

To identify the causal variants for the four INDELs regions, we constructed a rank scoring system based on the public functional databases. As shown in **Table 3**, we found that rs5777156, rs71450133 and rs34057993 were related to multiple regulatory elements (promoter histone marks, enhancer histone marks and TFBS) in multiple tissues or cell lines, while rs58404727 is located in a desert region. Furthermore, rs5777156 was located in the promoter histone marks and enhancer histone marks in the A549 EtOH 0.02pct lung carcinoma cell line in the ENCODE database; rs71450133 also showed enhancer histone marks in the A549 EtOH 0.02pct lung carcinoma cell line and in NHLF lung fibroblast primary cells in the ENCODE database. In the RegulomeDB annotation, the RegulomeDB score for rs5777156 was 3a, suggesting that rs5777156 might affect TF binding at the DNase peak. Meanwhile, rs71450133 may interact with the *VWF* and *CD9* genes through the 3D SNP annotation. The other two INDELs did not show any functional evidence in multiple databases. For these four new signals, we also identified seven candidate causal SNPs based on the rank scoring system (**Table S5**). At 1p13.2, a non-coding variant, rs12567622 in *MAGI3*, were predicted as the causal variant. At 4q28.2, the most plausible target SNP was rs72618844, which also showed an enhancer histone mark in the A549 EtOH 0.02pct lung carcinoma cell line. At 12p13.31, the predicted causal SNPs include rs7304688, which is located in the regulatory element site in A549 EtOH 0.02pct lung carcinoma cell line. At 14q22.3, the rs10483677 SNP was a predicted causal variant. Further studies will be required to determine whether these SNPs are truly causal variants for each locus.

The relationship between INDELs and SNPs

To understand the effects of the INDELs or SNPs on lung cancer risk, we

examined the relationship between the two types of variations from the same loci. In the four known loci, we found that most of the INDELs were in considerable linkage disequilibrium (LD) with previously reported risk SNPs (r^2: 0.5~1.0; **Table S6**). However, 5 INDELs in the HLA region did not show high LD with known risk SNPs ($r^2 < 0.1$). We performed a conditional analysis to determine whether those 5 INDELs exerted independent effects from known SNPs for each locus. INDEL rs145093187 showed an independent signal after adjusting the reported SNPs through Bonferroni correction (OR = 0.86, 95CI% = 0.81-0.91, conditional $P = 5.10\times10^{-8}$), while other the INDELs did not reach the suggestive threshold **(Table S7)**. For the new loci, the regional plots provide the LD relationship between the INDELs and SNPs at a 1 Mb window (**Figure 2**). We found that although INDELs showed a strong effect on lung cancer risk, there were still SNPs with high LD ($r^2 > 0.8$) showing a stronger effect. We also conducted conditional analysis on the INDELs and top SNPs in each locus. By adding the SNP with the lowest *P* value into the model for each locus, neither the INDEL nor the SNP showed a significant signal (**Table S7**). Meanwhile, we also performed conditional analyses on the four candidate causal SNPs and four new INDELs in each locus. When we added the candidate causal SNPs to the model, the INDELs showed stronger effects at the statistical level. In half of the four conditional analyses, the INDELs remained nominally significant ($P < 0.05$) (**Table S8**).

Discussion

In this study, we conducted a genome-wide meta-analysis with 23,202 cases and 19,048 controls to systematically explore the associations between INDELs and lung cancer risk. We identified 19 signals for lung cancer risk, and 4 of them were first reported in lung cancer.

INDEL rs5777156 is an insertion lying in the *MAGI3* intron at 1p13.2. *MAGI3* acts as a scaffolding protein at cell-cell junctions, regulating various cellular and signaling processes, such as the *Ras* signaling pathway and *PTEN* pathway. Previous studies showed that *MAGI3* could downregulate Wnt/β-catenin signaling,

suppressing malignant glioma cell phenotypes [28], and competes with *NHERF-2* to negatively regulate *LPA2* receptor signaling in colon cancer cells [29]. Additionally, INDEL rs5777156 and the predicted causal variant were all present in regulatory elements, including promoter and enhancer histone marks in a lung carcinoma cell line based on the ENCODE database, suggesting that rs5777156 may affect lung cancer risk through transcript regulation.

Our study also identified a new risk locus at 4q28.2 marked by INDEL rs58404727 mapping to 65 kb upstream of *RP11-184M15.2*, which is a lncRNA with little functional evidence. However, the predicted causal variant SNP rs72618844 showed promoter and enhancer histone marks in A549 lung carcinoma cell line. INDEL rs58404727 may be a tagging signal at this locus, while rs72618844 affects lung cancer risk.

INDEL rs71450133 is a deletion that maps to 23 kb upstream of *PLEKHG6* at 12p13.31. Genetic variants at 12p13.31 have been shown by previous studies to be associated with colorectal cancer risk in East Asians [30]. Although the function of *PLEKHG6* in tumors is unclear, some studies showed that *PLEKHG6* might regulate the invasion activity of breast cancer cells [31,32]. In the eQTL analyses, rs71450133 was associated with the expression of several genes, and 4 of them were tumor related. *GAPDH* encodes a member of the glyceraldehyde-3-phosphate dehydrogenase protein family and can interact with proteins participating in DNA repair [33]. *USP5*, namely ubiquitin specific peptidase 5, plays an important role in ubiquitination. *USP5* expression has been proven to be associated with several cancer types, such as hepatocellular carcinoma, glioblastoma and pancreatic cancer [34-36]. Previous studies have shown that *USP5* had many cellular targets and stabilizes multiple proteins, such as p53 [37]. *TPI1*, triosephosphate isomerase 1, encodes a crucial enzyme in the carbohydrate metabolism, and previous studies have shown its expression level might be associated with several cancer types [38,39]. Another gene, *MLF2* or Myeloid Leukemia Factor 2, is related to myeloid leukemia and leukemia, and *MLF2* knockdown may reduce tumor initiation and metastasis in breast cancer [40]. Functional annotation based on ENCODE suggested that

rs71450133 and its high LD SNPs are located in regulatory elements in A549 EtOH 0.02pct lung carcinoma cell line.

Another new susceptibility locus, 14q22.3, was marked by INDEL rs34057993, which is a deletion located in the intron of non-coding RNA *OTX2-AS1*, an *OTX2* antisense RNA at 14q22.3. *OTX2*, which encodes a member of the bicoid subfamily of homeodomain-containing transcription factors, has been implicated as a potential driver of medulloblastoma tumorigenesis [41,42]. Although rs34057993 and its LD SNPs did not show any promoter or enhancer histone marks, genes associated with INDEL rs34057993 were cancer-related, it is possible that rs34057993 may act by regulating the expression of genes to influence lung cancer risk.

In this study, we found four novel risk loci for lung cancer, as well as illustrated the relationships between INDELs and SNPs. In the reported regions, most of the significant INDELs were correlated with previously reported SNPs, especially in 5p15.33 and 15q25.1. In the HLA region, we found a novel signal that was independent of the previously reported SNPs. Considering the complex LD and haplotype structure in the HLA region [43], the novel INDEL may be a true association. In the new regions, we also observed INDELs that did not harbor the lowest *P* values and showed high LD with nearby SNPs. The effects of the INDELs were decreased after adjusting for the top SNP in each region. This suggests that the presented SNPs promote more stable effects in both known and new regions. However, it is generally assumed that SNPs with the most significant signal usually tag causal variants with a small effect. After conducting conditional analysis on seven potential causal SNPs, we found that the INDELs in the new loci were still nominally significant. Thus, it is possible that the INDELs may also be both causal and tagging variants. The combination of these variants with small effects together could lead to lung cancer. The functional annotation results confirmed our insights. In new region, two INDELs, rs5777156 and rs34057993, showed enhancer histone marks in regulatory regions, which may influence enhancer activity in lung cancer. Meanwhile, the most significant SNPs in those two regions did not show strong functional evidence. This means INDELs could also be a causal variant, which could regulate gene expression and affect the risk of lung cancer. The comprehensive

annotation of each locus also identified potential causal variants in high LD with the INDELs. Interestingly, we noticed that all 19 significant INDELs mapped to the non-coding region (intronic or intergenic region). INDELs in the coding region can result in frameshift and non-frameshift mutations, which are relatively severe mutations and more likely to be observed in Mendelian diseases or tumors [9,11]. Overall, the limitation of the present study is that we only evaluated the functional evidence from available databases for the identified INDELs, further functional experiments are needed to better understand INDEL mechanisms in lung cancer carcinogenesis.

In conclusion, we performed a large-scale case-control study to evaluate INDELs and their risk for lung cancer, and four new risk loci at 1p13.2, 4q28.2, 12p13.31 and 14q22.3 were identified. Our findings indicate that INDELs could be potentially functional genetic variants for lung cancer risk.

Supplemental Data

Supplemental Data include one figure and eight tables.

Acknowledgments

We thank the study participants and research staff for their contributions and commitment to this study. This work was supported by National Natural Science of China (81521004, 81820108028) and the Priority Academic Program for the Development of Jiangsu Higher Education Institutions [Public Health and Preventive Medicine] and Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (PPZY2015A067). This work was supported by the National Institutes of Health and the National Cancer Institute (CA209414 and CA092824 to D.C.C).

Declaration of Interests

The authors declare no competing interests.

Data availability

The INDEL data sets used during the current study are available at the database of Genotypes and Phenotypes (dbGaP) under accession phs001273.v1.p1 (TRICL-ILCCO OncoArray European data) and phs000336.v1.p1 (DCEG Lung Cancer Study).

Web Resources

OncoArray, http://epi.grants.cancer.gov/oncoarray/;
dbGap, https://www.ncbi.nlm.nih.gov/gap;
1000 Genomes Project, http://www.internationalgenome.org/;
GTEx, http://www.gtexportal.org/home/;
TCGA, https://cancergenome.nih.gov/;
ENCODE, https://www.encodeproject.org/;
3DSNP, http://cbportal.org/3dsnp/;
CADD, http://cadd.gs.washington.edu/home;
PINES, http://genetics.bwh.harvard.edu/pines;
RegulomeDB, http://www.regulomedb.org/;

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer 2015;136: E359-86.

2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: a cancer journal for clinicians 2015;65: 87-108.

3. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, Lan Q, Abnet CC, Amundadottir LT, Figueroa JD, Landi MT, Mirabello L, et al. Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. Journal of the National Cancer Institute 2015;107: djv279.

4. Dai J, Shen W, Wen W, Chang J, Wang T, Chen H, Jin G, Ma H, Wu C, Li L, Song F, Zeng Y, et al. Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population. International journal of cancer 2017;140: 329-36.

5. Bosse Y, Amos CI. A Decade of GWAS Results in Lung Cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2018;27: 363-79.

6. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome research 2006;16: 1182-90.

7. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Human molecular genetics 2010;19: R131-6.

8. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature 2015;526: 68-74.

9. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics 2014;133: 1-9.

10. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013;502: 333-9.

11. Ye K, Wang J, Jayasinghe R, Lameijer EW, McMichael JF, Ning J, McLellan MD, Xie M, Cao S, Yellapantula V, Huang KL, Scott A, et al. Systematic discovery of complex insertions and deletions in human cancers. Nature medicine 2016;22: 97-104.

12. Hoffmann TJ, Van Den Eeden SK, Sakoda LC, Jorgenson E, Habel LA, Graff RE, Passarelli MN, Cario CL, Emami NC, Chao CR, Ghai NR, Shan J, et al. A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer discovery 2015;5: 878-91.

13. Tao R, Hu S, Wang S, Zhou X, Zhang Q, Wang C, Zhao X, Zhou W, Zhang S, Li C, Zhao H, He Y, et al. Association between indel polymorphism in the promoter region of lncRNA GAS5 and the risk of hepatocellular carcinoma. Carcinogenesis 2015;36: 1136-43.

14. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, Byun J, Dunning A, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological

subtypes. Nature genetics 2017;49: 1126-32.

15. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. American journal of human genetics 2009;85: 679-91.

16. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, Yang L, Dai J, Hu L, Tan W, Li Z, Deng Q, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. Nature genetics 2011;43: 792-6.

17. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, Casey G, Hunter DJ, Sellers TA, Gruber SB, Dunning AM, Michailidou K, et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2017;26: 126-35.

18. Liu Q, Cirulli ET, Han Y, Yao S, Liu S, Zhu Q. Systematic assessment of imputation performance using the 1000 Genomes reference panels. Briefings in bioinformatics 2015;16: 549-62.

19. Cheng Y, Wang C, Zhu M, Dai J, Wang Y, Geng L, Li Z, Zhang J, Ma H, Jin G, Lin D, Hu Z, et al. Targeted sequencing of chromosome 15q25 identified novel variants associated with risk of lung cancer and smoking behavior in Chinese. Carcinogenesis 2017;38: 552-8.

20. Dong J, Cheng Y, Zhu M, Wen Y, Wang C, Wang Y, Geng L, Shen W, Liu J, Li Z, Zhang J, Ma H, et al. Fine mapping of chromosome 5p15.33 identifies novel lung cancer susceptibility loci in Han Chinese. International journal of cancer 2017;141: 447-56.

21. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489: 519-25.

22. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511: 543-50.

23. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology 2011;9: e1001046.

24. Lu Y, Quan C, Chen H, Bo X, Zhang C. 3DSNP: a database for linking human noncoding SNPs to their three-dimensional interacting genes. Nucleic acids research 2017;45: D643-D9.

25. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 2014;46: 310-5.

26. Corneliu A Bodea AAM, Heiko Runz, Shamil R Sunyaev. Phenotype-specific information improves prediction of functional impact for noncoding variants. bioRxiv 2016.

27. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. Genome research 2012;22: 1790-7.

28. Ma Q, Yang Y, Feng D, Zheng S, Meng R, Fa P, Zhao C, Liu H, Song R, Tao T, Yang L, Dai J, et al. MAGI3 negatively regulates Wnt/beta-catenin signaling and suppresses malignant phenotypes of glioma cells. Oncotarget 2015;6: 35851-65.

29. Lee SJ, Ritter SL, Zhang H, Shim H, Hall RA, Yun CC. MAGI-3 competes with NHERF-2 to negatively regulate LPA2 receptor signaling in colon cancer cells. Gastroenterology 2011;140: 924-34.

30. Zhang B, Jia WH, Matsuda K, Kweon SS, Matsuo K, Xiang YB, Shin A, Jee SH, Kim DH, Cai Q, Long J, Shi J, et al. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nature genetics 2014;46: 533-42.

31. Wu D, Asiedu M, Wei Q. Myosin-interacting guanine exchange factor (MyoGEF) regulates the invasion activity of MDA-MB-231 breast cancer cells through activation of RhoA and RhoC. Oncogene 2009;28: 2219-30.

32. Wu D, Haruta A, Wei Q. GIPC1 interacts with MyoGEF and promotes MDA-MB-231 breast cancer cell invasion. The Journal of biological chemistry 2010;285: 28643-50.

33. Kosova AA, Khodyreva SN, Lavrik OI. Role of Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) in DNA Repair. Biochemistry Biokhimiia 2017;82: 643-54.

34. Liu Y, Wang WM, Lu YF, Feng L, Li L, Pan MZ, Sun Y, Suen CW, Guo W, Pang JX, Zhang JF, Fu WM. Usp5 functions as an oncogene for stimulating tumorigenesis in hepatocellular carcinoma. Oncotarget 2017;8: 50655-64.

35. Izaguirre DI, Zhu W, Hai T, Cheung HC, Krahe R, Cote GJ. PTBP1-dependent regulation of USP5 alternative RNA splicing plays a role in glioblastoma tumorigenesis. Molecular carcinogenesis 2012;51: 895-906.

36. Li XY, Wu HY, Mao XF, Jiang LX, Wang YX. USP5 promotes tumorigenesis and progression of pancreatic cancer by stabilizing FoxM1 protein. Biochemical and biophysical research communications 2017;492: 48-54.

37. Dayal S, Sparks A, Jacob J, Allende-Vega N, Lane DP, Saville MK. Suppression of the deubiquitinating enzyme USP5 causes the accumulation of unanchored polyubiquitin and the activation of p53. The Journal of biological chemistry 2009;284: 5030-41.

38. Jiang H, Ma N, Shang Y, Zhou W, Chen T, Guan D, Li J, Wang J, Zhang E, Feng Y, Yin F, Yuan Y, et al. Triosephosphate isomerase 1 suppresses growth, migration and invasion of hepatocellular carcinoma cells. Biochemical and biophysical research communications 2017;482: 1048-53.

39. Linge A, Kennedy S, O'Flynn D, Beatty S, Moriarty P, Henry M, Clynes M, Larkin A, Meleady P. Differential expression of fourteen proteins between uveal melanoma from patients who subsequently developed distant metastases versus those who did Not. Investigative ophthalmology & visual science 2012;53: 4634-43.

40. Dave B, Granados-Principal S, Zhu R, Benz S, Rabizadeh S, Soon-Shiong P, Yu KD, Shao Z, Li X, Gilcrease M, Lai Z, Chen Y, et al. Targeting RPL39 and MLF2 reduces tumor initiation and metastasis in breast cancer by inhibiting nitric oxide synthase signaling. Proceedings of the National Academy of Sciences of the USA 2014;111: 8838-43.

41. Adamson DC, Shi Q, Wortham M, Northcott PA, Di C, Duncan CG, Li J, McLendon RE, Bigner DD, Taylor MD, Yan H. OTX2 is critical for the maintenance and progression of Shh-independent medulloblastomas. Cancer research 2010;70: 181-91.

42. Bunt J, Hasselt NE, Zwijnenburg DA, Hamdi M, Koster J, Versteeg R, Kool M. OTX2 directly activates cell cycle genes and inhibits differentiation in medulloblastoma cells. International journal of cancer 2012;131: E21-32.

43. de Bakker PI, Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. Human molecular genetics 2012;21: R29-36.

Tables

Table 1. The association between the INDELs in the new regions and lung cancer risk.

Chr.	INDEL	Gene	INS/DEL	INFO.	Major	Minor	EUR [a]	EAS [b]	Overall Results [c]		
									OR (95%CI)	*P*	Het *P*
1p13.2	rs5777156	*MAGI3*	Insertion	0.999	-	A	0.24	0.61	0.92 (0.89,0.95)	9.10×10^{-8}	0.837
4q28.2	rs58404727	*RP11-184M15.2*	Deletion	0.999	T	-	0.02	0.33	1.19 (1.11,1.28)	5.25×10^{-7}	0.191
12p13.31	rs71450133	*PLEKHG6*	Deletion	0.986	AA	-	0.18	0.38	1.09 (1.05,1.13)	8.83×10^{-7}	0.990
14q22.3	rs34057993	*OTX2-AS1*	Deletion	0.975	G	-	0.17	0.27	0.90 (0.87,0.94)	7.64×10^{-8}	0.587

a: The effect allele frequencies of the insertion or deletion in 1000 Genomes EUR samples;

b: The effect allele frequencies of the insertion or deletion in 1000 Genomes EAS samples;

c: The OR (95%CI) and *P* value for the meta-analysis were fixed-effects model;

INFO.: imputaion quality info.; Het *P*: *P* value for heterogeneity test.

Table 2. The association between the INDELs in the known regions and lung cancer risk.

Chr.	INDEL	Gene	INS/DEL	Major	Minor	EUR [a]	EAS [b]	Overall Results [c]		
								OR (95%CI)	*P*	Het *P*
5p15.33	rs34218850	*TERT*	Deletion	C	-	0.34	0.19	1.14 (1.11,1.18)	7.98×10^{-18}	0.097
6p21.32	rs200675567	*HLA-DQA1*	Deletion	C	-	0.11	0.16	0.90 (0.86,0.96)	4.03×10^{-7}	0.412
6p21.32	rs9279532	*NOTCH4*	Deletion	G	-	0.12	0.05	1.11 (1.07,0.94)	2.77×10^{-7}	0.588
6p21.33	rs550239034	*POU5F1*	Deletion	TT	-	0.25	0.46	0.92 (0.89,0.95	2.12×10^{-7}	0.055
6p21.33	rs549219764	*HCP5*	Deletion	G	-	0.20	0.02	1.11 (1.07,1.15)	2.01×10^{-9}	0.091
6p22.1	rs9280949	*RPP21*	Insertion	-	T	0.09	0.06	1.16 (1.11,1.22)	2.53×10^{-10}	0.468
6p22.1	rs139089584	*LINC00533*	Insertion	-	TTTG	0.29	0.54	0.92 (0.89,0.95)	2.40×10^{-7}	0.145
6p22.1	rs34832458	*HLA-G*	Insertion	-	T	0.39	0.24	0.92 (0.89,0.94)	1.29×10^{-9}	0.077
6p22.1	rs374787445	*HLA-F-AS1*	Deletion	C	-	0.18	0.23	1.11 (1.07,1.15)	6.56×10^{-9}	0.323
6p22.2	rs145093187	*BTN2A1*	Insertion	-	T	0.12	0.05	0.87 (0.83,0.91)	9.44×10^{-9}	0.478
11q23.3	rs139157129	*MPZL2*	Deletion	A	-	0.48	0.43	0.93 (0.90,0.95)	1.90×10^{-7}	0.864
15q25.1	rs577626090	*CHRNA5*	Deletion	AAAAG	-	0.37	0.03	1.29 (1.25,1.33)	9.91×10^{-64}	0.945
15q25.1	rs138784116	*CHRNB4*	Deletion	AGG	-	0.37	0.14	0.89 (0.86,0.92)	4.65×10^{-14}	0.655
15q25.1	rs143284856	*MORF4L1*	Insertion	-	TT	0.47	0.12	1.11 (1.08,1.14)	1.29×10^{-12}	0.732
15q25.1	rs61655864	*CHRNA5*	Deletion	A	-	0.29	0.77	0.81 (0.79,0.84)	6.24×10^{-37}	0.068

a: The effect allele frequencies of the insertion or deletion in 1000 Genomes EUR samples;

b: The effect allele frequencies of the insertion or deletion in 1000 Genomes EAS samples;

c: The OR(95%CI) and *P* value for the meta-analysis were fixed-effects model;

Het *P*, heterogeneity *P* value.

Table 3. Comprehensive functional annotations for the INDELs in the new regions.

Chr.	SNP	Region	INS/DEL	Gene	Enhancer [a]	Promoter [a]	TFBS [a]	3D Score [a]	3D Interaction Gene [a]	CADD [b]	RegulomeDB [c]	PINES [d]
1p13.2	rs5777156	Intronic	Insertion	*MAGI3*	6	1	2	2.300	-	4.264	3a	0.243
4q28.2	rs58404727	Intergenic	Deletion	*RP11-184M15.2*	0	0	0	1.820	-	3.743	7	0.499
12p13.31	rs71450133	Intergenic	Deletion	*PLEKHG6*	11	0	0	3.810	*VWF*, *CD9*	6.315	6	0.089
14q22.3	rs34057993	Intronic	Deletion	*OTX2-AS1*	18	1	0	6.960	-	1.310	7	0.061

a: Enhancer, promoter and TFBS were obtained from 3DSNP based on the ENCODE database. 3DSNP was the overall function score and the interacting gene reflected the three-dimensional interaction genes.

b: CADD was used to evaluate the relative deleteriousness.

c: RegulomeDB was used to identify DNA features and regulatory elements in non-coding regions in the human genome.

d: PINES provided a powerful in silico method to prioritize and finely map the functional non-coding variants. SNPs with lower P values indicated more abundant functions.

Figure Titles and Legends

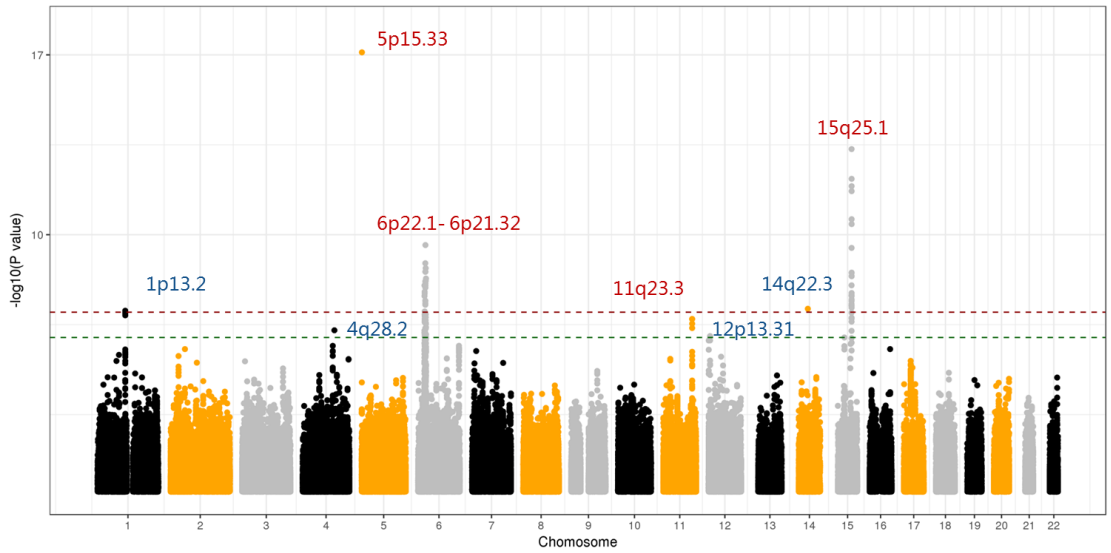

Figure 1. Manhattan plots of INDEL associations with lung cancer risk.
The x-axis represents the chromosomal location and the y-axis represents the -log10 (*P* value). Red, previously known loci and blue, new loci identified in this analysis. The red line denotes the Bonferroni correction significance ($P = 1.03 \times 10^{-7}$) and the green line denotes the suggestive significance ($P < 1.0 \times 10^{-6}$).

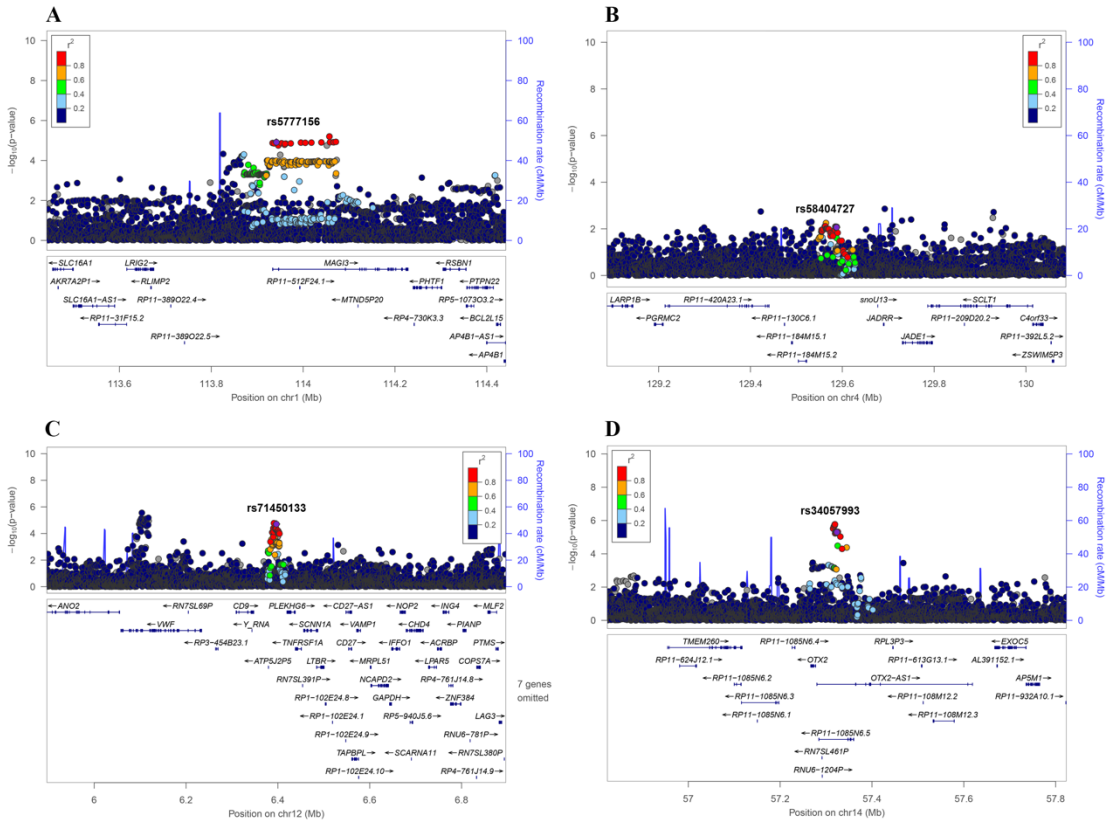

Figure 2. Regional plots of the 4 new regions, including (A) Chr1p13.2: rs5777156, (B) Chr4q28.2: rs58404727, (C) Chr12p13.31: rs71450133, and (D) Chr14q22.3: rs34057993.

The x-axis shows the chromosomal positions and the left y-axis shows the –log10 p values from an association test. The INDELs are shown as purple diamonds. The colors of the dots indicate the LD relationship between the most significantly associated INDELs and the remaining SNPs in the 500 kb region. The right y-axis shows the recombination rate between the SNPs. The genes within the region-of-interest are annotated with arrows indicating the direction of transcription.

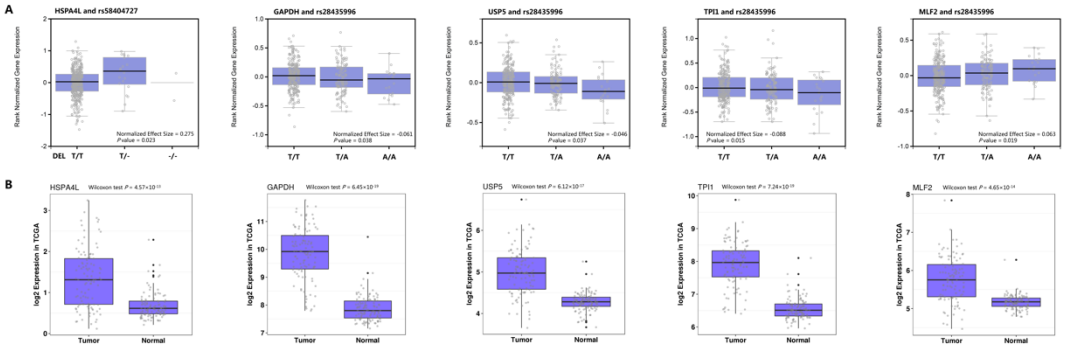

Figure 3. eQTL and differential expression of the INDELs among GTEx lung tissue and TCGA lung cancer data.